Temporal ordering of clinical events

Azad Dehghan

a.dehghan@manchester.ac.uk

The University of Manchester, School of Computer Science, Manchester, UK

Abstract

This report describes a minimalistic set of methods engineered to anchor clinical events onto a temporal space. Specifically, we describe methods to extract clinical events (e.g., Problems, Treatments and Tests), temporal expressions (i.e., time, date, duration, and frequency), and temporal links (e.g., Before, After, Overlap) between events and temporal entities. These methods are developed and validated using high quality datasets.

Keywords: Clinical event extraction, clinical named entity recognition, temporal information extraction, temporal relation extraction, temporal link identification, temporal expression recognition and normalisation, ner, tern, tlink.

1. Introduction

Temporal ordering of events from semi-/un-structured textual data (e.g., news article, clinical narrative) has important applications in many practical clinical applications such as questioning and answering (e.g., personal assistance), timeline analysis (e.g., event monitoring, pathway extraction), and text summarisation.

Chronological ordering of events involves the tasks of named entity recognition and classification (NER) or event extraction, including temporal entity recognition and normalisation (TERN), and temporal relation (TLINK) identification and classification.

Moreover, temporal ordering of events from textual clinical data include, at least, three NLP tasks: (1) event extraction (e.g., clinical events or concepts such as problems, treatments and tests), (2) temporal entity extraction: identification (e.g., 'January 4 1988', 'twice daily') and normalisation,

and (3) temporal relations extraction (*determine when a particular event occurred*).For example, in Figure 1 a number of events (highlighted) and TE (underlined) have been identified in a sample clinical narrative. Subsequently, the chronological ordering of events have been visualised in the given timeline.

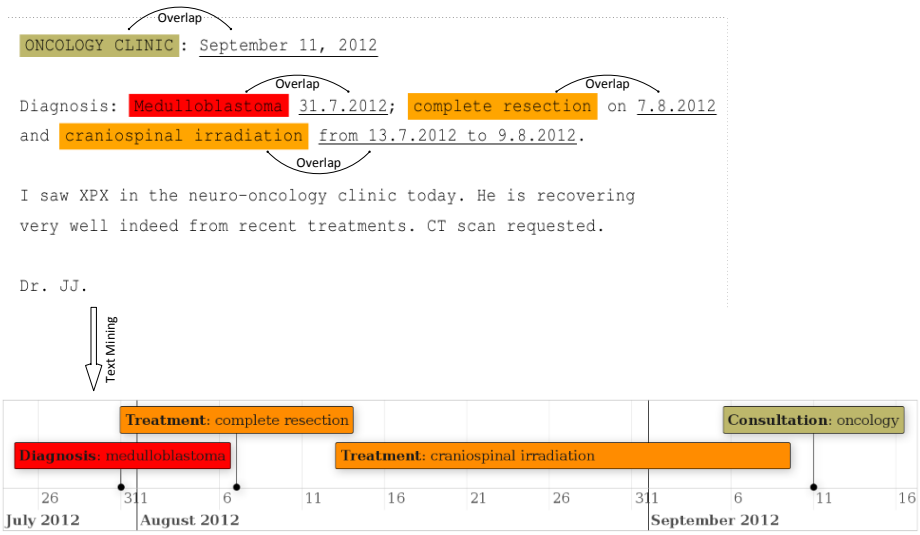

Figure 1: Chronologically ordered events from a sample clinical narrative

The methods described in this report has been inspired from a number of work derived from community held evaluations in relevant NLP tasks:

Event extraction

Recent work in clinical event extraction has been notably pushed by recent community held evaluation in clinical named entity recognition organised as part of 2010 [1] and 2012 [2] i2b2 challenges.

Temporal entity extraction

Likewise, temporal entity extraction has been notably pushed by a number of general domain SemEval/TempEval [3, 4, 5] and notably specific domain 2012 i2b2 [2] challenges.

Temporal relation extraction

The aim of temporal relation extraction is to anchor extracted events onto a temporal space. Recent work on this problem have resulted from the 2012 i2b2 [2] and more recently SemEval-2015 task 6, Clinical TempEval [6].

The remainder of this paper is structured as followed: Section 2 described the methods engineered to extract clinical events such as medical problems, treatments and tests. Section 3 describes the methods developed to identify and normalise (ISO-8601). Section 4 describes the temporal entity identification and classification approach. Section 5 presents the experiments, results and dicussions. The conclusion is given in the final Section 6.

This paper is largely self-contained

Note that this report is a reprint from the author's thesis [TBA] and significant improvement of intermediate results previously published [7].

A number of components described herein are available as open source[1]

2. Event extraction

The aim of the event extraction method is to identify broad clinical event categories such as, *Problem*, *Treatment* and *Test* and map them to a medical knowledge base such as the UMLS Metathesaurus for fine-grained semantic characterisation of event instances[2]. These event categories will collectively be referred to as EVENTs from henceforth. We have adopted the i2b2 definitions of concept or event categories which are largely based on the UMLS semantic types, but not limited by their coverage[3];

2.1. Methods

The core NER is a data-driven approach (using the state-of-the-art sequence labelling algorithm CRF) to identify clinical EVENTs from healthcare narratives.

[1] Clinical NERC http://sourceforge.net/projects/clinical-nerc/ and Clinical TERN http://sourceforge.net/projects/clinical-tern/

[2] No evaluation is provided on event/concept mapping.

[3] https://www.i2b2.org/NLP/Relations/assets/Concept%20Annotation%20Guideline.pdf

The EVENT extraction pipeline is made up of three main processing components: NLP pre-processing, the NER (see Figure 2), and Negation.:

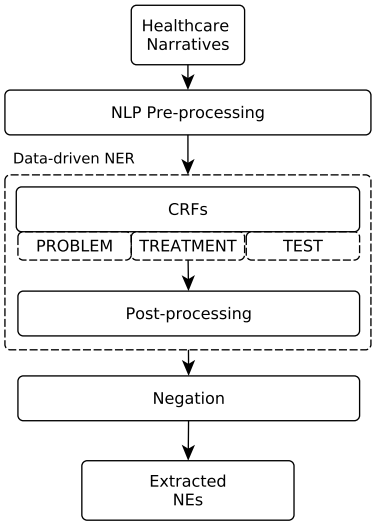

Figure 2: EVENT extraction architecture

The NLP pre-processing pipeline is made up of lexical and syntactic processing components, specifically: (1) Tokeniser, (2) sentence splitter, (3) word stemmer, (4) POS tagger, and (5) chunker / shallow parser.

Data-driven NER

The Data-driven NER component utilises separate CRFs trained for each EVENT category: *Problem*, *Treatment* and *Test*. A combination of the forward and backward feature selection approaches were adopted to select a total of 20 most discriminant features from an initial set of 120 features. The same set of features were used across all categories as our analysis showed this was the best fit. The extracted features can be clustered into two sets: lexical and syntactic, with four feature groups across (see the below list).

Lexical

- fg_1: the token string or alphanumeric character sequence
- fg_2: the stem of each token
- fg_3: the POS-tag for each token

Syntactic

- fg_4: the chunk tag for each token

Further, the feature space is also made up of contextual features of the neighbouring tokens with a feature window size of 5 or [-2,2] with respect to the current token. The *window size* corresponds to the number of tokens to the *left* and *right*, including the current token, of which contextual token features are considered. Specifically, for each token t and a given feature group fg, the feature space consists of: (t_{fg}), (t_{fg}+1), (t_{fg}+2), (t_{fg}-1), and (t_{fg}-2) (see Table 1).

Table 1: Feature template: clinical EVENTs

CRF feature template used for all EVENT categories: Problem, Treatment and Test.

fg_1**:Token**	fg_2**:Stem**	fg_3**:POS**	fg_4**:Chunk**
U00:%x[-2,1]	U05:%x[-2,2]	U10:%x[-2,3]	U15:%x[-2,4]
U01:%x[-1,1]	U06:%x[-1,2]	U11:%x[-1,3]	U16:%x[-1,4]
U02:%x[0,1]	U07:%x[0,2]	U12:%x[0,3]	U17:%x[0,4]
U03:%x[1,1]	U08:%x[1,2]	U13:%x[1,3]	U18:%x[1,4]
U04:%x[2,1]	U09:%x[2,2]	U14:%x[2,3]	U19:%x[2,4]

All CRFs were trained using a mix of BIO and W-BIO (W: single word, B: beginning, I: inside, O: outside) sequence label models with the following (default) CRF parameters: $C = 1.00$, ETA : 0.0001 and L2-regularisation algorithm.

The post-processing component contains three sub-components:

1. **Label fixer**
   This components corrects sequence label prediction from the NER component. These corrections are simple heuristics based on commonly observed errors in the training data set. Table 2 list the full set of heuristics utilised.
2. **Boundary adjustment**
   This component attempts to expand the event boundary by including contextual tokens to the right and left of predictions that possess POS/chunk tags that corresponded to nouns and noun phrases and their constituents including adjectives and determiners (e.g., 'a'; 'this';

Table 2: Label fixer heuristic

	Raw predictions	**Corrected predictions**
a	... O O O I I I I ...	... O O **B** I I I I ...
b	... O O O B O O O ...	... O O O B **I** O O ...
c	... O O O B O I I ...	... O O O B **I** I I...
d	... O O O B I I B I I ...	... O O O B I I **I** I I...

'her'). This sub-component is useful when the NER only tags part of an event. For example, if the NER component annotates the word 'severe', 'stomach', or 'ache' from the actual term 'severe stomach ache', this component would hypothetically capture the latter complete term boundary.

3. **False positive filter**
   This component removes common false positives predictions observed during the validation of the NER, i.e., common model prediction errors. Examples of false positives prediction include single character predictions (e.g., 'a'), pronouns (e.g., 'he'; 'she'), and determiners (e.g., 'the').

Negation

To identify negated clinical EVENTs we used the ConText negation tool as described in [8].

3. Temporal entity extraction

The TERN task involves the recognition and normalisation of TEs. TE are defined by TIMEX3 schema are grouped into four temporal types: *Date* (e.g., 'August 23, 1993'), *Time* (e.g., '2:23 p.m.'), *Frequency* (e.g., 'every morning'), and *Duration* (e.g., 'two weeks'). In addition, the *Date and time format: ISO-8601* standard is used to normalise TEs into a standardised format.

3.1. Methods

We propose a hybrid-based TER component, with a rule-based temporal normalisation component (ClinicalNorMA)[4]. The motivation for adopting a

[4]https://github.com/filannim/clinical-norma

hybrid approach for TER was to compare different approaches, and potentially combine the methods for the best possible performance.

Architecture

The TERN component is made up of the following components (see Figure 3 for a overview).

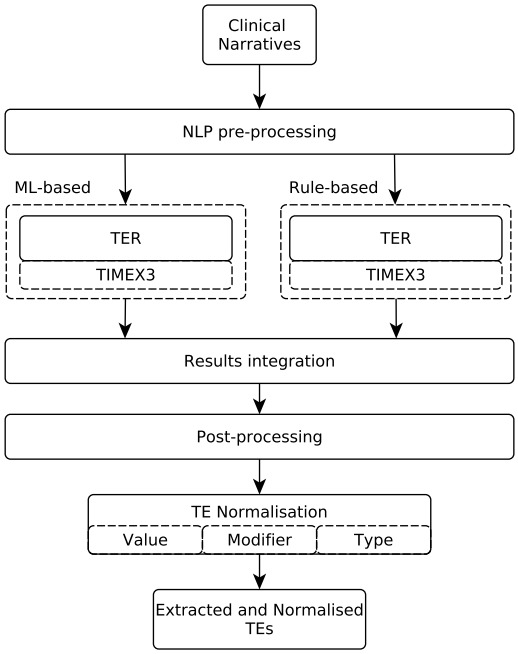

Figure 3: TERN architecture

A pre-processing pipeline is made up of the following NLP components: (1) tokeniser, (2) sentence splitter, and (3) semantic temporal resources.Specifically, several bespoke temporal knowledge resources were manual compiled and applied at this stage of processing to subsequently be utilised as features for the rule- and ML-based TER components. These semantic resources cover a broad set of temporal expression sub-categories:

- clinical frequency (e.g., qd (once a day), bid (twice a day))
- duration (e.g., 'over night', 'weekend', 'months')
- festival (e.g., 'Yom Kippur', 'Nowruz', 'Christmas')

- season (e.g., 'summer', 'spring', 'autumn', etc.)
- weekday (e.g., 'Monday', 'Tuesday', 'Wednesday', etc.)
- month (e.g., 'January', 'February', 'March', etc.)
- literal time (e.g., 'morning', 'afternoon', 'evening')
- temporal modifier (e.g., 'on', 'after', 'before')
- ordinal number (e.g., 'first', 'second', 'third', etc.)
- literal number (e.g., 'one', 'two', 'three', etc.)

Temporal expression recognition

The TER component consists of combined rule- and ML-based methods.

The rule-based component consist of a total of 65 rules containing patterns derived from an initial collocation extraction (i.e., bi- and tri-grams) and pattern analysis of TEs in the training data. For example, the TE patterns 'MM/DD/YYYY', 'MM/DD/YY', 'YYYY/DD/MM' and 'MM/DD' accounted for roughly 35% of temporal expressions in the training data (i2b2-TRC).

The rule set developed combines a few types of feature: (a) semantic: temporal categories derived from the set of specific temporal knowledge resources during the pre-processing (see previous sub-section), (b) lexical: such as common recurring expressions (e.g., 'postoperative day one', 'hospital day five', 'today'), and (c) pattern features e.g., 'MM/DD/YYYY', 'MM/DD/YY'.

The ML-based component was developed using a set of features selected based on an initial literature review, and further refinement using a combination of manual forward and backward feature selection approach. A total of 19 most discriminate features were selected from an initial set of 120 features. These features can be organised into three sets:

Lexical

- fg_1: the token string or alphanumeric character sequence
- fg_2: semantic temporal categories derived from the 'NLP pre-processing'

Orthographic

- fg_3: token kind given by the literal representation: *word*, *number*, *symbol*, or *punctiuation*
- fg_4: token-case given by the literal representation: *lower-case*, *upper-case*, *upper-initial*, *mixed-caps*, *all-caps*

Combined

- fg_5: concatenation of the features: fg_1, fg_2 and fg_4

In addition to these features, the feature space consists of contextual features. Specifically, we found that the optimal feature window size of 5 or [-2,2] was ideal for fg_1, fg_3 and fg_4, and a window size of 3 or [0,2] for fg_2 (Table 3 gives the complete feature space used).

Table 3: Feature template: clinical TER

CRF feature template used for the TER.

fg_1**:Token**	fg_2**:Dictionary**
U00:%x[-2,1]	
U01:%x[-1,1]	
U02:%x[0,1]	U05:%x[0,2]
U03:%x[1,1]	U06:%x[1,2]
U04:%x[2,1]	U07:%x[2,2]
fg_3**:TokenKind**	fg_4**:TokenCase**
U08:%x[-2,5]	U13:%x[-2,6]
U09:%x[-1,5]	U14:%x[-1,6]
U10:%x[0,5]	U15:%x[0,6]
U11:%x[1,5]	U16:%x[1,6]
U12:%x[2,5]	U17:%x[2,6]
fg_5**:Combined**	
U18:%x[0,1]/%x[0,2]/%x[0,4]	

The ML-based module uses a a state-of-the-art sequence labelling algorithm (CRF) trained with the IO token representation schema with the following (default) CRF parameters: $C = 1.00$, ETA : 0.0001 and L2-regularisation algorithm.

Results integration

The output of the ML- and rule-based methods are combined at the mention level: union of the respective overlapping and non-overlapping outputs.

Post-processing

A rule-based post-processing component was developed in order to correct obvious and systematic errors from the hybrid TER method. This component removes common false-positives predictions identified during the development of the TER component. Common examples include single character predictions and non-related but similar numerical expressions e.g., pulmonary artery pressure measures (e.g., '42/21') and other (partial) numerical expressions such as telephone, fax and ward numbers.

TE normalisation

The ClinicalNorMA [7] is adopted as the TE normalisation component. The normaliser is based on the general domain normalisation component TRIOS [9]. Further, ClinicalNorMA is rule-based and adheres to the TIMEX3 schema, specifically, the extended schema described in [10].

4. Temporal relation identification and classification

The aim of temporal relation extraction is the chronological ordering of events. The identification of temporal links between entity pairs such as EVENTs (e.g., *Problem*, *Treatment*, *Test*), TEs, and EVENTs and TEs, as well as the subsequent classification of these links into predefined categories (e.g., *After*, *Before*, *Overlap*) is known as TLINK extraction.

The TLINK method developed and described herein is rule-based. The developed approach is motivated by a gap in current literature of pure knowledge driven methods for clinical TLINKs extraction (see Section **??**).

The developed method has two main components. The first component takes as input extracted clinical concepts (*Problem*, *Treatment*, and *Test*) and TEs (*Date*, *Time*, *Duration* and *Frequency*), and generates TLINK candidate pairs (the *identification* step) and subsequently *classifies* the identified links into three different categories: *After*, *Before*, or *Overlap*. A final component derives the transitive closure (refer to Appendix 7) of relations extracted in order to generate implied relations that have been missed by the preceding TLINK method.

Figure 4 shows an abstract representation of the methodology. The remaining part of this section describes our methods in detail.

TLINK identification and classification

A notable difference between previous work and our approach is that we use (i) a pure rule-based method for TLINK extraction, and (ii) combine the

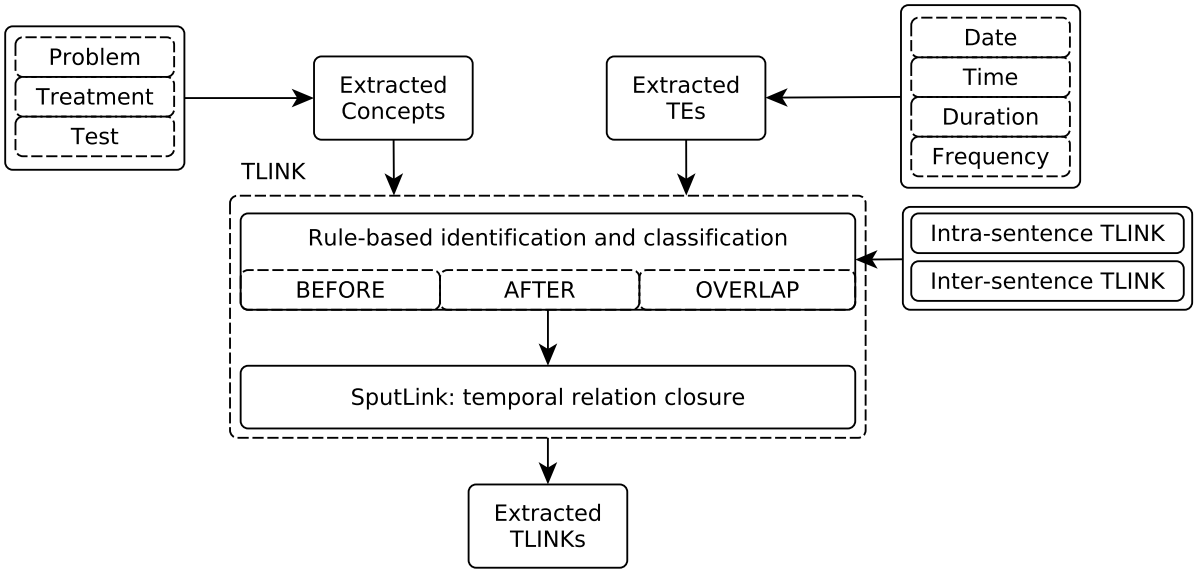

Figure 4: TLINK extraction architecture

TLINK candidate generation (identification) and classification into a single simultaneous component.

The rule based TLINK component is partitioned into two sub-components:

(1) intra-sentence: TLINKs within a sentence span;

(2) inter-sentence: TLINKs across sentences.

Intra-sentence TLINKs

In order to analyse intra-sentence TLINKs, we first performed an initial semi-automatic analysis in the development dataset. For each sentence containing a TLINK, the TLINK pairs were abstracted to their respective EVENT or TIMEX3 types. Additionally, any context to the right and left of the TLINKs were removed to easily spot patterns. Subsequently, the abstracted TLINK pairs were manually analysed for common patterns by the given TLINK category. For example, in the following sentences (a, b) the underlined EVENTs and TEs are part of six different TLINKs (or three TLINKs per sentence):

(a) 'The patient reported vomiting, nausea and headaches.'

(b) 'The patient received steroids for his swelling in 2006.'

In the following list, all pair-wise EVENTs or TE, part of TLINK is abstracted to their respective label and any context to the left and right of the pair-wise link is removed (illustrated by being ~~strikeout~~).

(a_1) '~~The patient reported~~ PROBLEM, PROBLEM ~~and headaches~~.

(a_2) '~~The patient reported vomiting,~~ PROBLEM and PROBLEM.'

(a_3) '~~The patient reported~~ PROBLEM, nausea and PROBLEM.'

(b_1) '~~The patient received~~ TREATMENT for PROBLEM ~~in 2006~~.'

(b_2) '~~The patient received steroids for~~ PROBLEM in DATE.'

(b_3) '~~The patient received~~ TREATMENT for his swellings in DATE.'

This approach enabled us to profile various TLINK categories and formalise extraction rules based on common abstraction patterns. For example, Table 4 lists a number of common patterns found and their typically associated TLINK category.

Profiling of TLINKs revealed the occurrence of different types of relations at the sentence level which we group into three different types: *co-ordinate*, *prepositional*, and *other* TLINKs. Further, these three types of TLINKs directly correspond to the type of extraction rules, which take advantage of corresponding features that characterised them. Specifically:

- **co-ordinate TLINKs** are links that are characterised by EVENTs separated by co-ordinate conjunctions such as 'and', 'or', or comma (i.e., ','). For example, in the sentence (a) above, all events are co-ordinate TLINKs. In the development dataset we observed that co-ordinate TLINKs as predominately categorised as 'overlap'.

- **prepositional TLINKs** are characterised by EVENTs/TEs that are linked by a prepositions. For example, in sentence (b), the preposition 'for' between the two EVENTs indicates the presence of a TLINK (in this particular example the TLINK is [TREATMENT] after [PROBLEM]).

Table 4: TLINK patterns

This table show common patterns semi-automatically extracted from the development/training dataset. The patterns listed in this tables make up the largest and most obvious TLINK patterns observed.

TLINK abstraction patterns		**Typical TLINK**
PROBLEM and PROBLEM	→	[PROBLEM] *Overlap* [PROBLEM
PROBLEM, PROBLEM	→	[PROBLEM] *Overlap* [PROBLEM]
TREATMENT on DATE	→	[TREATMENT] *Overlap* [DATE]
TREATMENT in DATE	→	[TREATMENT] *Overlap* [DAT]E
TREATMENT for PROBLEM	→	[TREATMENT] *Before* [PROBLEM]
TREATMENT of PROBLEM	→	[TREATMENT] *Before* [PROBLEM]
TEST showed PROBLEM	→	[TEST] *Before* [PROBLEM]
PROBLEM after TREATMENT	→	[PROBLEM] *After* [TREATMENT]
TREATMENT post TEST	→	[TREATMENT] *After* [TEST]

- **other TLINKs** are links that do not fit in either of the previously characterised types. A notable number of other TLINKs are characterised by linking verbs between EVENTs. For example, in the sentence 'The TEST *revealed* PROBLEM', TEST is linked, by the verb 'revealed', to PROBLEM (in this particular example the TLINK is: [TEST] *Before* [PROBLEM]).

Table 5 lists and describes the type of features used to extract intra-sentence TLINKs.

Inter-sentence TLINKs

TLINKs that span across sentences fall into two characterised types: SECTIME and co-referential TLINKs.

- **SECTIME TLINKs** represent the largest proportion of inter-sentence TLINKs (e.g., in the full i2b2-TRC corpus, 45.87% of all TLINKs are SECTIME links [10]). These are links anchored to relevant document section. Specifically, in the i2b2-TRC dataset, all events within 'History of Present Illness' or related sections are linked to the admission date, and events within the 'Hospital course' section are linked to dis-

charge date. SECTIME links are predominately categorised as *Before*.

Notably, it is not uncommon that clinical narratives do not contain sectime. More commonly events are anchored to the document creation time also known as DocTimeRel (document creation time relation).

- **Co-referential TLINKs** are EVENT co-references. These type of TLINKs are characterised as multiple EVENT mentions that refer to the same EVENT.

The approach for these two types of inter-sentence TLINKs differed. In the i2b2-TRC datasets, for development and testing, SECTIME TLINKs were addressed in a three step approach:

(1) extract admission and discharge dates;

(2) apply Section Boundary Detection (SBD), i.e., identify 'history of present illness' and 'hospital course' sections accordingly;

(3) anchor each EVENT in a given document section to the appropriate section date and set each TLINK category to *Before*.

However, in the case-study data there were couple notable differences to how SECTIME TLINKs were extracted. Namely, as there only existed one section time i.e., the DRD, the SBD was omitted and each EVENT was anchored to the DRD. In addition, while each TLINK category was initially set to the default link type *Before*, we observed a number of common events that occurred on the DRD: routine clinical measurements such as weight, height, blood pressure, and similar. These contained TLINK type were all amended accordingly to *Overlap*.

Co-referential TLINKs are approached by considering a novel feature: lexical-level similarity (i.e., comparing literal strings with no additional features considered) between EVENTs in a given clinical note. A combined token- and character-level string similarity metric SoftTFIDF algorithm [11] was adopted to determine the *similarity* between two candidate events. Specifically, the SoftTFIDF component take as input two strings and outputs a similarity score: a real number between [0,1]; where 1 is a perfect match and 0 the vice versa. The optimum threshold of 0.8 was determined through systematic experimentation with the i2b2-TRC development set.

The co-referential TLINK pseudo method developed is given below:

(1) using SoftTFIDF, $n^2 - 1$ comparisons are done between events in a given document (including across document sections, if any);

(2) if the SoftTFIDF similarity score between any pair-wise EVENTs is greater or equal to the threshold (0.8): create a TLINK between EVs with the link category: *Overlap*.

TLINKs features

Table 5 list the type of features used across both intra- and inter sentence TLINK methods. The features are used as part of formalised rules and heuristics to identify and classify TLINKs and include:

Table 5: TLINK extraction features

The features listed herein were used for both TLINK identification and classification; description of each feature type follows this table. Nota bene: EV=EVENT and ST=SECTIME.

Feature type	**Inter-sentence**			**Intra-sentence**	
	EV-EV	EV-TE	EV-ST	EV-EV	EV-TE
String similarity	✓				
Position information			✓	✓	✓
Distance information				✓	✓
Preposition				✓	✓
Conjunction				✓	✓
TE-related	✓				✓
NE-related	✓			✓	✓

Description of feature types listed in Table 5 follows.

- **String similarity**: specifically, string similarity score between pair-wise EVENTs derived from SoftTFIDF were used to extract co-referential TLINKs;
- **Position information**: the position of an EVENT within a given section (SECTIME TLINKs);
- **Distance information**: (i) token distance between entity pairs, and (ii) number of EVENT and TE between entity pairs;

- **Preposition**: between two candidate pairs e.g., 'in', 'on', 'after', 'before' and so forth;
- **Conjunction**: lexical cues between two candidate pairs e.g., 'and', 'both' and so forth;
- **TE-related**: TE type i.e., date, time, duration, and frequency;
- **EVENT-related**: EVENT information such as type i.e., *Problem*, *Treatment*, *Test*; including HrQoL concept categories) and negation information;

Temporal links closure

In order to capture implied TLINKs not captured by the initial rule-based method the transitive closure may be calculated. The final TLINK component engineered calculates the full set of transitive relations or temporal closure of links extracted using the initial rule-based component. Description of transitive closure is given in Appendix 7. However, the explicit results including transitive closure has not been included, except the inherit evaluation provided by the TempEval-3 metric.

5. Experiments, Results and Discussions

5.1. Data

The NER, TERN and TLINK methods presented in this report were developed and validated using a set of publicly available research datasets. The NLP research datasets used were obtained from the clinical TM challenges organised by the i2b2[5]. Specifically, these datasets are derived from the following shared-tasks:

(i) The 2010 i2b2 4^{th} Shared Task; referred to as *i2b2-CARC* hereafter [1], and

(ii) The 2012 i2b2 6^{th} Shared Task; referred to as *i2b2-TRC* hereafter [2].

Table 6 provides details such the size (number of documents across training and test datasets).

[5]the research datasets provided by i2b2 are not entirely public, and require data user agreements to be signed; https://www.i2b2.org/NLP/DataSets/.

Table 6: NLP datasets

This table shows the NLP datasets used in this report.

Dataset	Annotation	Training	Test
i2b2-TRC	EVENT[6],TIMEX3,TLINK	190	120
i2b2-CARC	EVENT[7]	170	256

These datasets were produced using multiple annotators, including domain experts. Specifically, the i2b2-TRC was produced using eight expert annotators, four of whom had medical background; the i2b2-CARC was produced using twelve annotators including six with medical background[8].

EVENT

The dataset utilised to engineer the event extraction method was composed of the i2b2-TRC and i2b2-CARC corpora. A total of 736 discharge summaries was used across the training (616 documents) and test (120 documents; i2b2-TRC test dataset). Table 7 shows the label distribution by event/concept category across the combined datasets used in this report.

Table 7: EVENT label distribution

In this report, the i2b2-TRC (training) and i2b2-CARC (training and test) data was combined as the training dataset, while the i2b2-TRC test dataset was used as the held-out test data for the clinical NER method described herein.

EVENT	Training	Test
Problem	24,330	4,309
Treatment	17,773	3,285
Test	16,062	2,173
Total	58,165	9,767

Table 8 show the IAA for i2b2-TRC [2, p.808][9] and Table 9 show the IAA

[8] Annotation task information regarding i2b2-CARC corpus was obtained by email from responsible researcher Brett South, Senior scientist (currently) at University of Utah, Department of Biomedical Informatics.

[9] These statistics are computed for *Problem*, *Treatment* and *Test*

for i2b2-CARC dataset[10]. The IAA scores confirm that recognition of event boundaries for both i2b2-TRC and i2b2-CARC is a fairly straight forward task for manual processing; with the identification of *Problem*, *Treatment* and *Test* event boundaries being a simpler task (see Table 9). Likewise, classification of EVENT *type* and concept negation reveal to be a relatively straight forward manual annotation task for appropriately trained experts.

Table 8: i2b2-TRC: EVENT IAA

EVENT	*Avg.P&R*	κ
Span (strict)	0.83	-
Span (lenient)	0.87	-
Type	0.93	0.90
Negation	0.97	0.21

Table 9: i2b2-CARC: EVENT IAA

EVENT	*Avg.P&R*
Span (strict)	0.85
Span (lenient)	0.91

[10]These statistics are computed across six different EVENTs: *Problem*, *Treatment*, *Test*, *Occurrence*, *Evidential* and *Clinical department*. Only the first three EVENT categories are considered in this report.

TIMEX3

The i2b2-TRC dataset was used for the development and evaluation of the TERN component. A total of 310 discharge summaries was used across the development (190 documents) and test (120 documents) datasets. Table 10 and Table 11 show the label distribution across the dataset by TE type and the IAA, respectively [2, p.808]. Notably, while the IAA shows a fairly good agreement for the recognition of TE spans (with strict boundary identification proving more challenging), normalisation of TE (i.e., *value*) seems even more challenging for manual efforts.

Table 10: TIMEX3 label distribution

Type	**Training**	**Test**
Date	1,641	1,222
Duration	407	341
Frequency	249	197
Time	69	60
Total	2,366	1,820

Table 11: i2b2-TRC: TIMEX3 IAA

TIMEX3	$Avg.P\&R$	κ
Span (strict)	0.73	-
Span (lenient)	0.89	-
Type	0.90	0.37
Value	0.75	-
Modifier	0.83	0.21

TLINK

The temporal relation component was developed and validated using the i2b2-TRC dataset. Note that only TLINKs that include EVENTs such as *Problem*, *Treatment*, Test and TIMEX3 have been considered. Table 12 and Table 13 show the label distribution and the IAAs, respectively [2, p.808]. Notably, and comparably (i.e., versus EVENT and TE recognition tasks), it is apparent that TLINK identification is a challenging task (i.e., 0.39 in average precision-recall) for humans. However, manual effort for TLINK classification (i.e., *type*) show reasonable performance.

Table 12: TLINK label distribution

Type	**Training**	**Test**
Before	11,981	10,488
Overlap	7,276	5,694
After	1,415	1,275
Total	20,672	17,457

Table 13: i2b2-TRC: TLINK IAA

TLINK	$Avg.P\&R$	κ
Span (strict)	0.39	-
Span (lenient)	-	-
Type	0.79	0.3

5.2. Event extraction

We explored a number of sequence label models: IO, BIO and W-BIO (where, W: single token word; B: beginning; I: inside; O: outside) in addition to a set of post-processing components. For the development/validation experiments we used the training data (Table 7) which we split into a validation training set (500 documents) and a validation test set (116 documents).

Table 14: EVENT extraction validation test results

The validation test set results are obtained by training the models on a set of 500 documents and testing on a validation test set of 116 (shown here). The best results, horizontally or by EVENT category, are highlighted. From all models experimented, the IO model performed worst overall concept types, with strict scores being notably lower than other models (approximately 5% across all concept categories). Further, the difference between BIO and W-BIO were minimal: the BIO models permed slightly better for the Problem and Treatment categories while W-BIO performed better on identifying the Test category.

EVENT	**Model**	**Precision %** strict|lenient	**Recall %** strict|lenient	F_1**-measure %** strict|lenient
Problem	IO	67.46|84.33	70.22|**87.78**	68.81|86.02
	BIO	**73.20**|**85.95**	**74.63**|87.62	**73.91**|**86.78**
	W-BIO	72.32|85.83	73.54|87.28	72.92|86.55
Treatment	IO	73.63|89.36	70.65|**85.74**	72.11|87.51
	BIO	**79.45**|90.37	**74.70**|84.97	**77.00**|**87.59**
	W-BIO	79.41|**90.91**	73.45|84.09	76.31|87.37
Test	IO	75.00|89.20	72.13|**85.79**	73.54|87.47
	BIO	80.14|89.82	76.37|85.59	78.21|87.65
	W-BIO	**80.72**|**90.34**	**76.50**|85.62	**78.56**|**87.92**
Micro score	IO	71.31|87.13	70.88|**86.60**	71.09|86.86
	BIO	**76.92**|88.30	**75.13**|86.24	**76.02**|**87.26**
	W-BIO	76.67|**88.54**	74.33|85.84	75.48|87.17

Our experiments showed that the IO models performed consistently worst compared to BIO and W-BIO, with the latter two models showing little difference (see Table 14). For example, considering strict evaluation metrics, there is minimal difference between BIO and W-BIO models, while a notable

difference can be observed between IO and BIO/W-BIO models (approximately 5% micro F_1-measure). This suggests that BIO and W-BIO models are better suited for strict boundary identification of clinical concepts investigated compared to the IO sequence label schema. Moreover, when considering lenient evaluation scores, there is a minimal difference among all models, however, BIO and W-BIO models score consistently higher precision and F_1-measure while IO models score consistently higher recall.

The final evaluation or test results are presented in Table 15. These are consistent with our findings during validation (Table 14). As may be seen from both the validation and evaluation results, there is no notable difference between BIO and W-BIO models, except for W-BIO (*Test*) which shows notably better results.

In light of evaluation results that are comparable to IAA (Table [8,9]), we have omitted detailed error analysis.

Table 15: EVENT extraction results on the held-out test set
The results on the held-out test set showed similar trend to the validation results; the IO models have been excluded due to notably poor performance on the validation set. Further, similar to the validation results, BIO performed better on Problem and Treatment categories while W-BIO model performed best on the Test category.

EVENT	**Model**	**Precision %** strict|lenient	**Recall %** strict|lenient	**F_1-measure %** strict|lenient
Problem	BIO	81.52|90.68	82.62|**92.90**	82.07|91.29
	W-BIO	**81.91**|**90.84**	**82.80**|91.83	**82.35**|**91.33**
Treatment	BIO	87.24|94.43	**80.12**|**86.73**	83.53|**90.42**
	W-BIO	**88.00**|**94.72**	**80.12**|86.24	**83.88**|90.28
Test	BIO	85.48|93.02	82.88|90.20	84.16|91.59
	W-BIO	**86.45**|**93.49**	**83.71**|**90.52**	**85.06**|**91.98**
Micro score	BIO	84.22|92.39	81.84|**89.78**	83.01|91.07
	W-BIO	**84.85**|**92.66**	**82.10**|89.66	**83.45**|**91.13**

The final models selected for the clinical NER pipeline was BIO for *Problem* and *Treatment*, and W-BIO for *Test*. The final evaluation scores, including negation is given in Table 16

Table 16: The clinical NER performance

	F_1-**micro** % strict|lenient	**Negation**	**Negation** κ
EVENT	83.21|91.17	0.93	0.65

Impact analysis

In order to justify the use of various features, datasets, and post-processing components, a series of impact analysis have been conducted and shown in Table 17 (which shows the feature impact of different CRF features used), Table 18 (impact of datasets on the overall performance) and Table 19 (impact of post-processing components).

Table 17 shows the feature impact analysis by the micro score of EVENTs; lexical features have been used as the baseline. Notably, word stem have the most impact (+3% strict and +2% lenient F_1); POS and chunk features showed minimal impact on their own with the latter having a negative impact of -0.01% lenient F_1. Further, while POS and chunk features adversely affect the precision, both show a positive effect on recall.

Table 17: EVENT recognition: feature impact analysis
This table shows the feature impact of several CRF feature groups.

Feature vector	**EVENTs**		
	P-micro % strict|lenient	R-micro % strict|lenient	F_1-micro % strict|lenient
Baseline (Lexical)	82.56|92.34	76.79|85.88	79.57|88.99
Baseline + Stem	84.56|**92.96**	81.37|89.44	82.93|91.17
Baseline + POS	82.51|92.08	77.66|86.67	80.01|89.29
Baseline + Chunk	82.39|92.07	77.03|86.09	79.62|88.98
All features	**84.43**|92.50	**82.02|89.85**	**83.21|91.17**

Notably, using the i2b2-CARC corpus improved (strict and lenient) micro F_1-score with +17% and +12% (see Table 18).

Table 19 shows the impact of the post-processing sub-components. For example, while the label-fixer has a adverse effect on the precision (-5%

Table 18: EVENT recognition: dataset impact
This table shows the impact of the different datasets used to train the CRF models.

Dataset	**EVENTs**		
	P-micro % strict|lenient	R-micro % strict|lenient	F_1-micro % strict|lenient
i2b2-TRC	69.03|82.97	63.04|75.77	65.90|79.20
i2b2-TRC+i2b2-CARC	**84.43|92.50**	**82.02|89.85**	**83.21|91.17**

strict and -4% lenient), it has a positive impact on recall (+3% strict and +5% lenient). In addition, the label-fixer shows less than -1% (strict) and more than +1% (lenient) impact on the F_1-score. Boundary adjustment showed a positive effect on all strict metrics, and expectedly with no effect on lenient scores. The FP filter showed a slight positive impact on precision, and interestingly vice-versa on recall.

Table 19: EVENT recognition: post-processing impact analysis
This table lists the performance impact of the various post-processing components.

Component	**EVENTs**		
	P-micro % strict|lenient	R-micro % strict|lenient	F_1-micro % strict|lenient
No post-processing	88.09|96.06	77.64|84.66	82.54|90.00
Only label-fixer	82.85|92.23	80.81|**89.97**	81.82|91.09
Only boundary-adjustment	**89.34**|96.06	78.73|84.66	**83.70**|90.00
Only FP filter	89.14|**96.45**	77.63|84.00	82.99|89.79
All post-processing	84.43|92.50	**82.02**|89.85	83.21|**91.17**

5.3. Temporal entity extraction

We explored a number of methods in order to adopt the best approach for TER (validation results are given in Table 20). For the ML-based method, we experimented with various sequence label schemas (i.e., IO, BIO and W-BIO). Notably, we discovered that all sequence label models explored performed relatively similar in terms of lenient scores, but W-BIO and BIO mod-

els performed notably better in terms of strict scores (e.g., 3-4% F_1-measure). However, the strict rule-based method outperformed all ML models both in terms of lenient and strict scores (over 90% lenient F_1-score).

Table 20: TER validation results

*The ML-based component was validated on the i2b2-TRC training data which was split 60/40% for training and validation respectively. *The rule-based results shown was obtained using the whole training data.*

Method	**Precision %** strict|lenient	**Recall %** strict|lenient	**F_1-measure %** strict|lenient
IO	66.03|86.94	67.17|88.44	66.60|87.69
BIO	71.26|87.85	70.95|87.47	71.10|87.66
W-BIO	71.80|87.85	71.49|87.47	71.65|87.66
Rule-based*	**78.66**|**89.64**	**80.15**|**91.34**	**79.40**|**90.48**

Using the official i2b2-TRC test set, we further evaluated the various ML models (using the complete training set to derive the models) and the rule-based method. In addition, we explored a number of combination the various ML models and the rule-based method (results are given in Table 21).

Table 21: TER evaluation on the held-out test set

This table shows the evaluation results of various ML-, rule- and hybrid-based methods on the official i2b2-TRC test.

Method	**Precision %** strict|lenient	**Recall %** strict|lenient	**F_1-measure %** strict|lenient
IO	64.42|87.10	66.65|90.11	65.51|88.58
BIO	67.45|86.63	69.56|89.34	68.49|87.96
W-BIO	68.22|86.47	68.41|86.70	68.31|86.58
Rule-based	**77.29**|**89.64**	76.65|88.90	**76.97**|89.27
IO+Rule-based	72.03|86.62	**78.41**|**94.29**	75.09|**90.29**
BIO+Rule-based	71.15|86.05	77.64|93.90	74.25|89.81
W-BIO+Rule-based	71.66|85.73	78.08|93.41	74.73|89.40

The evaluation on the held-out test set (Table 21) shows a similar trend to the validation results (Table 20) in terms of strict scores i.e., W-BIO and

BIO performs notably better than IO: approximately +3%. This indicates good generalisable methods. However, the IO model shows more notable improvement (than the validation results) in terms of F_1-measure over the W-BIO (+2%) and BIO (0.62%) models. The rule-based methods performs better than all ML models, except for the IO model's lenient recall.

We also explored a number of combinations between various ML models and the rule-based method (union of the output of each respective method) in order to discover any possible improvements. In particular, since the normalisation of TE is more important than recognition results, we are specifically interested in improved recall. The combination of the IO model and the rule-based method showed the best overall performance. A notable improvement, in terms of lenient recall, of +4.18% and +5.39% compared the best ML model (IO) and the rule-based method respectively, was achieved by the 'IO+rule-based' method. Similarly, the strict recall was improved with +8.85% and +1.76% over the best ML model (BIO) and the rule-based method respectively. In addition, the best F_1-measure of 90.29% was also achieved with the 'IO+rule-based' method. As expected, the rule-based method achieved the best precision of all methods. This slightly exceeds the state-of-the-art system [12], which reported an F_1-score of 90.03%.

The normalisation scores reproduced using the i2b2-TRC test dataset are given in Table 22. As apparent by the 'primary score' TERN task is a challenging task and an open research problem.

Table 22: TE normalisation results

This table gives the normalisation scores. The primary score is the product of the TER lenient F_1-measure and normalisation value accuracy and is considered the main TERN metric.

Type	Value	Modifier	Primary score
0.8473	0.7044	0.8275	0.63

While automated recognition of TEs have shown comparable and exceeding human-level benchmark results (e.g., [2, 5]), normalisation remain a challenge—both for human and automated methods. For instance, the current state-of-the-art clinical TERN methods achieve a mere 66% (primary score) which is just lower than the human benchmark of 66.75% [2]. Similarly, the state-of-the-art system [12] achieved a 73% accuracy for the normalised value

attribute, notably lower to the human benchmark of 75%. Regardless, these scores, either automated or human, are notably lower than common IE score of +90% which is typically considered as good.

One of the notable challenges of TERN is the normalisation of relative expressions (e.g., 'two weeks ago' 'post-operative day').

5.4. Temporal relation extraction

Evaluation metrics

The methods described herein have been validated using multiple evaluation methods/metrics. The main evaluation metric used in the 2012 i2b2 temporal relation challenge [10] was TempEval-3 type evaluation metrics where the 'reduced graph' or redundant relations (i.e., a relation is redundant if it can be inferred from other relations) are removed. The *TempEval-3 evaluation metric* used is described below:

- Precision: the total number of reduced system output TLINKs that can be verified in the gold standard closure divided by the total number of reduced system output TLINKs.
- Recall: the total number reduced gold standard output TLINKs that can be verified in the system closure divided by the total number of reduced gold standard output TLINKs.

We initially developed and evaluated our method using gold standard EVENTs and TEs; the results of these experiments are shown in Table 23 and Table 24. In addition, an end-to-end evaluation where EVENTs, TEs and TLINKs are all tagged using bespoke methods (described in Sections [2,3,4] respectively) is shown in Table 26.

As expected, fairly precision-bias results were achieved, as that was the aim during design and development. This leaves room for future work to further extend the current method in order to balance recall and to further optimise the overall score.

A direct comparison cannot be made between our results and work on the full i2b2-TRC dataset [10] due to the reason that our experiments were based on a reduced set of TLINKs. The full i2b2-TRC dataset included pairwise TLINKs of six different EVENTs, three more than used in our experiments. We did not include *Occurrence*, *Evidential* and *Clinical department* as they were not relevant/useful for characterising patient journeys.

Table 23: TLINK development set results

This table shows the performance of the TLINK pipeline on the development/training dataset. We used two evaluation metrics: common precision-recall and the TempEval-3.

Evaluation setting	**Precision %**	**Recall %**	**F_1-measure %**
Customary precision-recall	81.40	55.06	65.69
TempEval-3 precision-recall	80.43	48.05	62.59

Nonetheless, we note the performance of the best systems evaluated on the full i2b2-TRC dataset as a point of reference. The best systems to-date, using gold annotations (for clinical EVENTs and TEs) achieved a F_1-measure of 69% [13, 14]. As previously mentioned in Chapter **??**, both [13] and [14] use complex hybrid methods with rule based components for candidate generation (i.e., TLINK identification). For classification of TLINKs, [13] uses a combination of CRF and SVM, whilst [14] use a combination of MaxEnt and SVM for TLINK classification. In contrast, our method uses a knowledge based approach to recognise and simultaneously classify TLINKs. Our approach achieved an overall score of 62.96% F_1-measure on the held-out test set (Table 24). Further, considering common IE evaluation metrics, where system predictions are evaluated against manually annotated gold dataset without any further manipulation of labels, our approach achieved 65.34% with customary and 62.96% F_1-measure using TempEval-3 metrics.

Table 24: TLINK results on the held-out test set

This table shows the results of the TLINK pipeline on the held-out test set. The results are presented using common precision-recall and the TempEval-3 evaluation metric.

Evaluation setting	**Precision %**	**Recall %**	**F_1-measure %**
Customary precision-recall	81.51	54.52	65.34
TempEval-3 precision-recall	80.23	49.10	62.96

A comparison of results between the development (Table 23) and held-out test data (Table 24), indicate good generalisability of the methods developed. For instance, consider the minimal variation in F_1-measures between the development and test set. Except a small drop in F_1-score when not including temporal closure ('Regular (no closure)'), the results on the test dataset were slightly better than those on the development set.

Table 25 shows the specific component-based evaluation of SECTIME, intra-sentence and inter-sentence TLINKs. For each component, the held-out test set has been reduced to only the relevant type of TLINKs (i.e., when evaluating SECTIME, only SECTIME links are retained). These evaluation results are obtained using the test dataset with gold annotations.

Similar to the findings of the TLINK challenge described in [10], we found that SECTIME TLINKs were easiest to extract (see Table 25). Secondly, as expected, intra-sentence TLINKs where easier to extract than inter-sentence TLINKs (when exluding SECTIME TLINKs). Lastly, as concluded by previous work [10], and equally applicable to our rule based approach, a better method to generate candidate pairs would be beneficial to optimise recall.

Table 25: TLINK component based evaluation

This table shows the individual TLINK component based evaluation of the three TLINK sub-components: SECTIME, intra-sentence and inter-sentence TLINK methods. For each TLINK component evaluated the data has been reduced to only the relevant type of links.

TLINK	**Precision %**	**Recall %**	**F_1-measure %**
SECTIME	93.93	92.04	92.97
Inter-sentence	55.72	20.40	29.87
Intra-sentence	39.47	22.50	28.66

The component-based analysis also reinforces the conclusion that an extension of our method for recognition of candidate pairs ought to be explored. Currently, only neighbouring candidate EVENTs and co-referential inter-sentence TLINK are addressed. Extensions may include intra-sentence TLINKs that have multiple token distance in-between (e.g., first and last EVENTs in a sentence) and non co-referential inter-sentence TLINKs.

Moreover, Table 25 also shows the source of errors. Despite the aim of generating high precision rules, yet, it was challenging to replicate the manual

effort. However, the highly inconsistent annotations (i.e., IAA: 0.39) indicate that the TLINK annotation themselves were a notable source of generated errors.

End-to-end evaluation

Table 26 shows the end-to-end evaluation: all entities are derived from bespoke methods such as clinical NER (described in Chapter **??**), and the TERN method described in this chapter.

Table 26: TLINK end-to-end results on the held-out test set

This table shows the results of the end-to-end system output: all annotations are derived from the bespoke clinical NER, TERN and TLINK methods described in this report thus far.

Evaluation method	**Precision %**	**Recall %**	**F_1-measure %**
Customary precision-recall	78.27	48.21	59.67
TempEval-3 precision-recall	76.87	43.05	55.19

As a point of reference, [13] achieved 62.78% (F_1-measure) on the full i2b2-TRC dataset using the TempEval-3 evaluation method. Our method achieved 55.19% using the same metric on the reduced dataset (in terms of event categories considered). Further, evaluating our method as per typical IE evaluation (i.e., against the gold set without any manipulation to the temporal graph) we achieved a F_1-measure of 59.67%.

While our methods shows good precision, an apparent limitation is the recall. We hypothesis that a better approach to candidate generation can address the latter gap.

6. Conclusion

This report describes a set of NLP methods to order clinical events onto a temporal space or timeline. A number of notable observation were made from the validation of these methods:

- EVENTs or broad clinical concept categories (i.e., *Problem*, *Treatment*, *Test*) can be automatically extracted (using CRF) with comparable scores to human benchmark.

- negation of concepts can be automatically determined (using ConText negation tool with minor 'tailoring') with comparable accuracy to human benchmark.
- temporal entity identification can be automatically extracted with comparable score to human benchmarks.
- temporal entity normalisation is comparably challenging (even for humans). Further, determining the value (ISO-8601) was harder than type identification.
- TLINK extraction is overall an open research problem. DocTimeRel or SECTIME links can be extracted with good scores (93%) while intra- and inter-sentence links are notably more challenging to extract.

In future work, we will investigate the expansion of lexical features by incorporating lexical variant generation for EVENT extraction. The expansion of the TE normalisation component is currently being achieved. Additionally, expanding candidate generation heuristics and integrating machine learning classifiers are currently being investigated to improve the TLINK component.

7. Apppendice

Appendix A: Event extraction

Event conceptualisation

Table 27 shows the semantic definition of relevant event categories.

Table 27: Definition of EVENT categories

The definition of event categories are described according to the annotation guidelines

EVENT	Semantic type	Semantic group
Problem	acquired abnormality anatomical abnormality cell or molecular dysfunction congenital abnormality disease or syndrome injury or poisoning mental or behavioural dysfunction neoplastic process pathologic functions sign or symptom	Disorders
	bacterium virus	Living Beings
Treatment	antibiotic biomedical or dental material clinical drug pharmacologic substance steroid	Chemicals & Drugs
	drug delivery device medical device	Devices
	therapeutic or preventive procedure	Procedures
Test	diagnostic procedure laboratory procedure	Procedures

Appendix B: Transitive closure: an example

Transitive closure of a given relations computes all implicit relations or take into account its transitivity. Further, we can define a transitive relation as:

$$\forall_{a,b,c} \subseteq X : (aRb \wedge bRc) \Rightarrow aRc \tag{1}$$

For example, in Table 28 the letters A, B, C represent different EVENTs, and the arrows '$\rightarrow$', '$\leftarrow$', and '$\leftrightarrow$' represent the temporal relations 'before', 'after', and 'overlap' respectively. Hence, given the TLINKs: EVENT A before EVENT B, EVENT B after EVENT C, and EVENT A overlap EVENT C are represent as followed $A \rightarrow B$, $B \leftarrow C$, and $A \leftrightarrow C$ respectively.

Table 28: Transitive relations
This table shows a number of example of transitive relations.

If A $\rightarrow$ B and B $\rightarrow$ C, then A $\rightarrow$ C
If A $\leftarrow$ B and B $\leftarrow$ C, then A $\leftarrow$ C
If A $\leftrightarrow$ B and B $\leftrightarrow$ C, then A $\leftrightarrow$ C
If A $\rightarrow$ B and B $\leftrightarrow$ C, then A $\rightarrow$ C
If A $\rightarrow$ B and A $\leftrightarrow$ C, then C $\rightarrow$ B

References

[1] O. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, Journal of the American Medical Informatics Association 18 (5) (2011) 552–556. doi:10.1136/amiajnl-2011-000203.
URL http://dx.doi.org/10.1136/amiajnl-2011-000203

[2] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, Journal of the American Medical Informatics Association 20 (5) (2013) 806–813. doi:10.1136/amiajnl-2013-001628.
URL http://dx.doi.org/10.1136/amiajnl-2013-001628

[3] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, J. Pustejovsky, SemEval-2007 Task 15: TempEval Temporal Relation Identification, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 75–80.
URL http://www.aclweb.org/anthology/S/S07/S07-1014

[4] M. Verhagen, R. Saurí, T. Caselli, J. Pustejovsky, SemEval-2010 task 13: TempEval-2, in: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 57–62.
URL http://portal.acm.org/citation.cfm?id=1859674

[5] N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, J. Pustejovsky, SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events, and Temporal Relations (2013).

[6] S. Bethard, L. Derczynski, J. Pustejovsky, M. Verhagen, SemEval-2015 Task 6: Clinical TempEval, in: 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, 2015.

[7] A. Kovacevic, A. Dehghan, M. Filannino, J. A. Keane, G. Nenadic, Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives., Journal of the American Medical Informatics Association : JAMIA 20 (5) (2013) 859–66.
URL http://www.ncbi.nlm.nih.gov/pubmed/23605114

[8] H. Harkema, J. N. Dowling, T. Thornblade, W. W. Chapman, ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports, Journal of Biomedical Informatics 42 (5) (2009) 839–851. doi:http://dx.doi.org/10.1016/j.jbi.2009.05.002.
URL http://www.sciencedirect.com/science/article/pii/S1532046409000744

[9] N. UzZaman, J. Allen, TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 276–283.
URL http://www.aclweb.org/anthology/S10-1062

[10] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge., Journal of the American Medical Informatics Association : JAMIA 20 (5) (2013) 806–13. doi:10.1136/amiajnl-2013-001628.
URL http://www.ncbi.nlm.nih.gov/pubmed/23564629

[11] W. W. Cohen, P. Ravikumar, S. E. Fienberg, A comparison of string distance metrics for name-matching tasks (2003).
URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.3605

[12] S. Sohn, K. B. Wagholikar, D. Li, S. R. Jonnalagadda, C. Tao, R. Komandur Elayavilli, H. Liu, Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification, Journal of the American Medical Informatics Association 20 (5) (2013) 836–842. doi:10.1136/amiajnl-2013-001622.
URL http://dx.doi.org/10.1136/amiajnl-2013-001622

[13] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, H. Xu, A hybrid system for temporal information extraction from clinical text., Journal of the American Medical Informatics Association : JAMIA 20 (5) (2013) 828–35. doi:10.1136/amiajnl-2013-001635.
URL http://dx.doi.org/10.1136/amiajnl-2013-001635http://www.ncbi.nlm.nih.gov/pubmed/23571849

[14] C. Cherry, X. Zhu, J. Martin, B. de Bruijn, À la Recherche du Temps Perdu: extracting temporal relations from medical text in the 2012 i2b2

NLP challenge, Journal of the American Medical Informatics Association 20 (5) (2013) 843–848. doi:10.1136/amiajnl-2013-001624. URL http://dx.doi.org/10.1136/amiajnl-2013-001624